Adaptive Monte Carlo for Binary Regression with Many Regressors

D. Lamnisos*, J. E. Griffin† and M. F. J. Steel*

Abstract

This article describes a method for efficient posterior simulation for Bayesian variable selection in probit regression models with many regressors but few observations. A proposal on model space is described which contains a tuneable parameter. An adaptive approach to choosing this tuning parameter is described which allows automatic, efficient computation in these models. The methods is applied to the analysis of gene expression data.

1 INTRODUCTION

There are many problems that amount to the selection of a few variables, from a much larger set, with the aim of discriminating between two classes. For example, measurements of gene expression may be taken in a microarry experiment with the goal of finding a small subset of the genes that allow discrimination between two conditions such as disease or non-diseased, or two types of cancer. The standard approach compares the gene expression levels for the two groups on a gene-by-gene basis (see *e.g.* Dudoit *et al*, 2002). In this paper the problem is seen as variable selection in a probit regression model in a Bayesian framework (see *e.g.* Lee *et al* (2003), Sha *et al* (2003, 2004) and Yeung *et al* (2005)). This approach takes into account correlations between the variables and provides a measure of uncertainty in the choice of variables which also extends to any predictions. A Markov chain Monte Carlo (MCMC) method is used for posterior inference and our goal is to automatically produce a Markov chain with good mixing properties which gives accurate answers with the smallest possible run length. Alternative methods of estimation have been proposed. Variational methods for binary regression models with Gaussian priors have been discussed by *e.g.* Jaakola and Jordan (2000). Variable selection using sparsity-inducing priors have been discussed by *e.g.* Qi and Jaakola (2007).

Markov chain Monte Carlo (MCMC) has become the main tool for simulating from a target distribution, $\pi(x)$, which in our case will be a posterior distribution. One of the simplest such methods is the Random Walk Metropolis-Hastings (RWM) sampler, which generates a Markov chain whose marginal distribution is $\pi(x)$. Suppose that x_n is the current value of the chain, a potential new value of the chain, x', is proposed as

*Department of Statistics, University of Warwick, Coventry, CV4 7AL, U.K. and † Institute of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, CT2 7NF, U.K. Correspondence to M. Steel, Email: M.F.Steel@stats.warwick.ac.uk, Tel.: +44(0)24-76523369, Fax: +44(0)24-76524532

$x' = x_n + \epsilon_n$ where ϵ_n is a random perturbation drawn from some distribution (often taken to be normal) with mean zero and standard deviation σ. The potential value is accepted as the next value of the chain $x_{n+1} = x'$ or rejected $x_{n+1} = x_n$ with a probability given by the Metropolis-Hastings acceptance ratio. This method is popular because careful tuning of σ leads to the optimal RWM sampler and it can be combined with Gibbs sampling to produce simulation schemes for a wide-range of models. For many target distributions, the optimal RWM sampler occurs when the average acceptance probability $\bar{\tau}$ is 0.234 (see *e.g.* Roberts and Rosenthal, 2001 and Sherlock and Roberts, 2009). Tuning of the algorithm is simple since $\bar{\tau}$ will typically fall as the standard deviation σ is increased. The average acceptance rate $\bar{\tau}$ can be estimated from pilot runs with a fixed value of σ and the standard deviation can be adjusted so that the estimated $\bar{\tau}$ is close to 0.234.

Recently, there has been interest in adaptive Monte Carlo methods where the distribution of the proposal, x', is adjusted during the MCMC run. These methods are difficult to implement in general since the Markov property is violated and standard theory for convergence of the chain to the target distribution does not apply. However, convergence to the target distribution can be verified for particular forms of adjustment. The first adaptive algorithm that could be shown to converge to the target distribution was introduced by Haario *et al* (2001) who used methods from Stochastic Approximation. This important idea and other methods are reviewed by Andrieu and Thoms (2008).

This paper is concerned with extending these methods to problems in variable selection with many regressors. The posterior distribution will be complicated and vast (there will be 2^p potential models if there are p potential regressors). Adaptive methods are important because MCMC methods often mix slowly (and so proposals that encourage good mixing are important) and the running of many pilot runs is unsatisfactory due to the large number of iterations needed to give good estimates of posterior summaries. Such methods have been previously applied to variable selection problems by Nott and Kohn (2005). They allow the probability that a particular variable is proposed to be included in or removed from the model to adapt over the chain. This is rather different to the method developed here where the number of variables, rather than which variables, is adapted over the chain. The method uses a form of proposal for variable selection described by Lamnisos *et al* (2009) which can be tuned in a similar way to an RWM sampler.

The paper is organised as follows: Section 2 describes the Bayesian approach to variable selection and some MCMC algorithms for posterior exploration, Section 3 describes a tuneable proposal for variable selection problems, Section 4 describes adaptive versions of the algorithms, Section 5 includes some numerical examples that demonstrate the utility of the approach and Section 6 includes some concluding comments.

2 BAYESIAN VARIABLE SELECTION

Let $y_1, y_2, \ldots, y_n$ be observations taking values 0 and 1 where the i-th observation is associated with a set of p regressors $x_{i1}, x_{i2}, \ldots, x_{ip}$. In this paper, it is assumed that p is much larger than n. For example, y_i could be the disease status of the i-th patient and $x_{i1}, x_{i2}, \ldots, x_{ip}$ is a vector of gene expression measurements or proteomic measurements (see *e.g.* Lee *et al* (2003), Sha *et al* (2003, 2004) and Yeung *et al* (2005)). It is assumed that only a subset of the regressors are needed to predict y_i and we use the indicator

variables $\gamma_1, \gamma_2, \ldots, \gamma_p$ to represent whether the i-th regressor is included in the model ($\gamma_i = 1$ if a regressor is included and $\gamma_i = 0$ otherwise). Let $p_\gamma = \sum_{i=1}^{p} \gamma_i$ be the size of model γ. It is assumed that observations are generated by the probit model:

$$p(y_i = 1) = \Phi(\alpha + x_i^\gamma \beta_\gamma),$$

where Φ is the cumulative distribution function of the standard normal distribution, x_i^γ is a vector containing only the elements of x_i for which $\gamma_i = 1$ and β_γ is a p_γ-dimensional vector of regression coefficients We also define X_γ to be the submatrix of X which only includes those columns for which $\gamma_i = 1$. Within a Bayesian analysis, both β_γ and γ are given prior distributions. The regression coefficients are assumed to be $\beta_\gamma|\gamma \sim \mathrm{N}(0, V_\gamma)$ where V_γ might follow the so-called g-prior form $V_\gamma = c(X_\gamma' X_\gamma)^{-1}$ or the independent form $V_\gamma = cI$ where I is the identity matrix and c is a positive scalar. The prior assumes that $p(\gamma_i = 1) = w$ which implies that the number of included variables is binomially distributed with mean nw and variance $nw(1-w)$. The hyperparameter w can be interpreted as the prior probability of including a variable in the model.

The posterior distribution of γ and $\theta_\gamma = (\alpha, \beta_\gamma)$ is not available analytically and so computational methods are needed. Markov chain Monte Carlo methods are a popular class of algorithms to fit the model. Several methods have been proposed in the literature. Holmes and Held (2006) describe one algorithm, which uses data augmentation with a vector z based on a representation described in Albert and Chib (1993):

1. Generate $z = (z_1, z_2, \ldots, z_n)'$ where $z_i \sim \mathrm{N}(\alpha + x_i^\gamma \beta_\gamma, 1)$ where $z_i > 0$ if $y_i = 1$ or $z_i < 0$ if $y_i = 0$.
2. Select model γ' with probability $q(\gamma'|\gamma)$.
3. Jump to the model γ' with probability
$$\alpha(\gamma, \gamma') = \min\left\{1, \frac{\pi(\gamma')q(\gamma|\gamma')\pi(z|\gamma')}{\pi(\gamma)q(\gamma'|\gamma\pi(z|\gamma)}\right\}.$$
4. If the jump is accepted, draw a sample $\theta_{\gamma'} \sim \mathrm{N}((X_{\gamma'}^{\mathrm{T}} X_{\gamma'} + V_{\gamma'}^{-1})^{-1} X_{\gamma'}^{\mathrm{T}} z, (X_{\gamma'}^{\mathrm{T}} X_{\gamma'} + V_{\gamma'}^{-1})^{-1})$.

Alternatively, one can update γ and θ_γ jointly. Green (2003) describes an Automatic Generic Sampler which is extended by Lamnisos *et al* (2009).

1. Draw a sample $\theta_\gamma \sim \mathrm{N}((X_\gamma^{\mathrm{T}} X_\gamma + V_\gamma^{-1})^{-1} X_\gamma^{\mathrm{T}} z, (X_\gamma^{\mathrm{T}} X_\gamma + V_\gamma^{-1})^{-1})$.
2. Select model γ' with probability $q(\gamma'|\gamma)$.
3. Propose $\theta_{\gamma'}$ in the following way. Let μ_γ and Σ_γ be an approximation of the mean and variance of the posterior distribution of θ_γ and let B_γ be the Cholesky decomposition of Σ_γ and $v = B_\gamma^{-1}(\theta_\gamma - \mu_\gamma)$. Then we propose $\theta_{\gamma'} = \mu_{\gamma'} + B_{\gamma'} v'$ where
$$v' = \begin{cases} (v_1, \ldots, v_{p_{\gamma'}})^{\mathrm{T}} & \text{if } p_{\gamma'} < p_\gamma \\ v & \text{if } p_{\gamma'} = p_\gamma \\ (v^{\mathrm{T}}, \epsilon^{\mathrm{T}})^{\mathrm{T}} & \text{if } p_{\gamma'} > p_\gamma \end{cases}$$
where $\epsilon = (\epsilon_1, \ldots, \epsilon_{p_{\gamma'} - p_\gamma})^{\mathrm{T}}$ has i.i.d. $\mathrm{N}(0,1)$ elements.

4. Jump to the model γ' and parameter $\theta_{\gamma'}$ with probability $\alpha(\gamma, \gamma', \theta_\gamma, \theta_{\gamma'}) =$

$$\min\left\{1, \frac{\pi(\gamma', \theta_{\gamma'})q(\gamma|\gamma')\pi(y|\theta_{\gamma'})|B_{\gamma'}|}{\pi(\gamma, \theta)q(\gamma'|\gamma)\pi(y|\theta_\gamma)|B_\gamma|} \times K\right\} \tag{1}$$

where

$$K = \begin{cases} (2\pi)^{-\frac{1}{2}(p_\gamma - p_{\gamma'})} \exp\left\{-\frac{1}{2}(\epsilon')^{\mathrm{T}}(\epsilon')\right\} & \text{if } p_{\gamma'} < p_\gamma \\ 1 & \text{if } p_{\gamma'} = p_\gamma \\ (2\pi)^{-\frac{1}{2}(p_\gamma - p_{\gamma'})} \exp\left\{\frac{1}{2}\epsilon^{\mathrm{T}}\epsilon\right\} & \text{if } p_{\gamma'} > p_\gamma, \end{cases}$$

and ϵ' is the obvious counterpart of ϵ.

5. Generate $z_1, z_2, \ldots, z_n$ where $z_i \sim \mathrm{N}(\alpha + x_i^\gamma \beta_\gamma, 1)$ where $z_i > 0$ if $y_i = 1$ or $z_i < 0$ if $y_i = 0$.

There are several methods of finding μ_γ and Σ_γ. Two are considered in this paper: the Laplace approximation and the Iterated Weighted Least Squares (IWLS) approximation for one iteration (see Lamnisos *et al* (2009) for more details).

Alternative automatic methods for moving between models are described by Brooks *et al* (2003) which are applied in this context by Lamnisos *et al* (2009). In the case of proposals that increase the model size, the coefficient vector is completed with $u_\gamma(\upsilon) = \mu + \sigma\upsilon$ where υ has a standard normal distribution. The Conditional Maximization method chooses μ to maximize the posterior distribution $\pi((\theta_\gamma, u_\gamma)|y)$ with respect to u_γ. The variance, σ^2, is chosen to ensure that the acceptance probability of the move with $u_\gamma(\upsilon) = \mu$ is 1 and is given by

$$\sigma = \left(\frac{\pi(y|\theta_\gamma, \gamma)\ \pi(\gamma)\ q(\gamma'|\gamma)\ c^{(p_{\gamma'} - p_\gamma)/2} \exp\frac{\mu^{\mathrm{T}}\mu}{2c}}{\pi(y|(\theta_\gamma, \mu), \gamma')\pi(\gamma')q(\gamma|\gamma')}\right)^{\frac{1}{p_{\gamma'} - p_\gamma}}.$$

Here

$$\alpha(\gamma, \gamma', \theta_\gamma, \theta_{\gamma'}) = \min\left\{1, \frac{\pi(y|\theta_{\gamma'}, \gamma')\ \pi(\gamma')\ q(\gamma|\gamma')}{\pi(y|\theta_\gamma, \gamma)\pi(\gamma)q(\gamma'|\gamma)} B\right\} \tag{2}$$

where

$$B = \exp\left\{\frac{1}{2}\left[\upsilon^{\mathrm{T}}\upsilon - (\mu + \sigma v)^{\mathrm{T}}(\mu + \sigma v)/c\right]\right\}\left(\frac{\sigma^2}{c}\right)^{\frac{p_{\gamma'} - p_\gamma}{2}}.$$

The pseudocode representation of the Conditional Maximization method is as follows: If at iteration t the current state is $(\theta_\gamma^{(t)}, \gamma)$, then

1. Draw a sample $\theta_\gamma \sim \mathrm{N}((X_\gamma^{\mathrm{T}} X_\gamma + V_\gamma^{-1})^{-1} X_\gamma^{\mathrm{T}} z, (X_\gamma^{\mathrm{T}} X_\gamma + V_\gamma^{-1})^{-1})$.
2. Select model γ' with probability $q(\gamma'|\gamma)$.
3. Determine the location μ and the scale σ of the proposal random variable u_γ as described above.
4. Generate $u_\gamma \sim N_{p_{\gamma'} - p_\gamma}(\mu, \sigma^2 I_{p_{\gamma'} - p_\gamma})$.

5. Set $\theta_{\gamma'} = (\theta_{\gamma}^{(t)}, u_{\gamma})$.

6. Jump to the model γ' and set $\theta_{\gamma'}^{(t+1)} = \theta_{\gamma'}$ with probability given by (2). Otherwise set $\theta_{\gamma}^{(t+1)} = \theta_{\gamma}^{(t)}$.

7. Generate $z_1, z_2, \ldots, z_n$ where $z_i \sim \mathrm{N}(\alpha + x_i^{\gamma}\beta_{\gamma}, 1)$ where $z_i > 0$ if $y_i = 1$ or $z_i < 0$ if $y_i = 0$.

3 A TUNEABLE PROPOSAL ON MODEL SPACE

Lamnisos *et al* (2009) propose a new general model proposal $q_{\zeta}(\gamma'|\gamma)$ which draws a new model in the following way:

1. A value $N^{(t)}$ is generated from a Binomial distribution with $N-1$ trials and success probability ζ.

2. One of three possible moves: Add, Delete and Swap is chosen uniformly at random. If Add is selected then $N^{(t)}+1$ regressors are chosen to be added to those included in γ to form γ', if Delete is selected then $N^{(t)}+1$ regressors are chosen to be removed from the model and if Swap is selected then $N^{(t)}+1$ regressors are swapped without changing the model size (provided $p_{\gamma} \geq N^{(t)}+1$; if not, the Add step is chosen).

This model proposal combines local moves with more global ones by changing simultaneously a block of variables. Two parameters determine this proposal: N is the maximum number of variables that can be changed from the current model γ and ζ determines the degree of "localness" since the mean number of variables changed is $1-\zeta+N\zeta$. The value of N will usually be fixed and the parameter ζ chosen to control the mixing of the chain. The application of this proposal to microarray data by Lamnisos *et al* (2009) suggests that the optimum effective sample size is obtained when the average acceptance rate fall in the range 0.25 to 0.40. This is true for a wide-range of sampling schemes. Rather like RWM samplers, this optimal choice of acceptance rate can be achieved by carefully tuning the parameter ζ of the model proposal using a series of pilot runs. In each pilot run, the sampler is run for a chosen value of ζ and the average acceptance rate calculated. If the acceptance rate is too high then ζ is increased in the next run and if the acceptance rate is too low then ζ is decreased in the next run. However, this tuning process is typically a computationally expensive task since trial and error is required.

As an alternative solution, we consider adaptive MCMC algorithms which can automatically handle this parameter tuning. The problem is similar to adaptation in RWM samplers since there is a tuneable proposal and an optimal acceptance rate to be achieved. We will extend the Adaptive Random Walk Metropolis (ARWM) algorithm proposed by Atchadé & Rosenthal (2005) to the variable selection problem. The ARWM algorithm automatically finds the optimal scale parameter in the RWM algorithm that results in the optimal acceptance rate $\bar{\tau} = 0.234$. Therefore, our adaptive MCMC algorithms automatically find ζ such that the resulting acceptance rate falls in the range 0.25 to 0.4. A comprehensive survey of recent advances in adaptive MCMC methodology and their applications can be found in Rosenthal (2008).

4 THE ADAPTIVE ALGORITHM

The ARWM algorithm of Atchadé & Rosenthal (2005) sequentially adapts the scale parameter σ of the RWM with a d-variate normal proposal density centred on the current value and with variance $\sigma^2 I$ and for which the stationary distribution is the positive continuous density $\pi(x)$. The entire past of the stochastic process is used to adapt the scale parameter σ. The resulting sequence of scale parameters $\left\{\sigma^{(t)} : t \in \mathbb{N}\right\}$ converges to an optimal value that leads to the optimal acceptance rate $\bar{\tau} = 0.234$. Atchadé & Rosenthal (2005) fix values ε_1 and A_1 such that $0 < \varepsilon_1 < A_1$ and defined the set $\Delta = \{\sigma : \varepsilon_1 \leq \sigma \leq A_1\}$. They also assume that Δ contains a unique value σ_{opt} which results in the optimal acceptance rate $\bar{\tau}$. Then, they define the following function of σ

$$\rho(\sigma) = \begin{cases} \varepsilon_1 & \text{if} \quad \sigma < \varepsilon_1 \\ \sigma & \text{if} \quad \sigma \in \Delta \\ A_1 & \text{if} \quad \sigma > A_1. \end{cases} \tag{3}$$

The aim of this function is to contain the adaptive algorithm inside Δ. Finally, they define a discount factor $s^{(t)}$, which is a positive sequence of real numbers such that $s^{(t)} = O(t^{-\lambda})$ for some constant $1/2 < \lambda <= 1$. This assumption ensures the ergodicity of the ARWM algorithm. The simple choice $s^{(t)} = a\, t^{-1}$ for some $a > 0$ will meet this condition. The pseudocode of the ARWM algorithm proceeds as follows:
If at iteration t the current state is $x^{(t)} \in \mathbb{R}^d$ and the scale parameter of the proposal density is $\sigma^{(t)} \in \Delta$, then

1. Generate $y \sim \mathrm{N}(x^{(t)}, (\sigma^{(t)})^2 I)$.
2. Set $x^{(t+1)} = y$ with probability

$$\alpha(x^{(t)}, y) = \min\left\{1, \frac{\pi(y)}{\pi(x^{(t)})}\right\},$$

   otherwise take $x^{(t+1)} = x^{(t)}$.
3. Compute

$$\sigma^{(t+1)} = \rho(\sigma^{(t)} + s^{(t)}(\alpha(x^{(t)}, y) - \bar{\tau})). \tag{4}$$

The acceptance rate is monitored by (4). The algorithm decreases the scale parameter $\sigma^{(t+1)}$ when the acceptance rate is small and increases $\sigma^{(t+1)}$ when the acceptance rate is high. Atchadé & Rosenthal (2005) showed under certain assumptions that the generated stochastic process is ergodic with stationary distribution the positive continuous density $\pi(x)$.

Turning to the MCMC algorithms of Section 2 with model proposal $q(\gamma'|\gamma)$, the parameter ζ behaves like the scale parameter σ of the RWM because values of ζ close to 0 yield more local moves and high acceptance rate and values of ζ close to 1 more global moves and small acceptance rate. Moreover, our applications to some gene expression datasets suggest an optimal acceptance rate $\bar{\tau}$ between 0.25 and 0.4. Therefore, we adopt ideas of the ARWM to develop adaptive version of each transdimensional MCMC sampler

described in Section 2. In our case $\zeta \in [0,1]$, thus values of ε_1 and A_1 are chosen close to 0 and 1, respectively. In our applications of the adaptive algorithms, ε_1 and A_1 are set to 0.01 and 0.99, respectively. The parameter ζ can be made adaptive by updating it at the t-th iteration in the following way, analogous to the RWM,

$$\zeta^{(t+1)} = \rho(\zeta^{(t)} + s^{(t)}(\alpha^{(t)} - \bar{\tau})) \tag{5}$$

where $\alpha^{(t)}$ is the acceptance probability at the t-th iteration of the chain.

All the algorithms of Section 2 can be made adaptive by updating ζ at each iteration using the recursion in (5). The pseudocode representation of the adaptive Holmes and Held algorithm adjusts the model proposal step and adds an extra step (step 4 below) in the corresponding non-adaptive algorithm. This pseudocode representation is as follows: If at iteration t the current state is $(\theta_\gamma^{(t)}, \gamma)$ and $\zeta^{(t)} \in [0,1]$, then

1. Generate $z_1, z_2, \ldots, z_n$ where $z_i \sim \mathrm{N}(\alpha + x_i^\gamma \beta_\gamma, 1)$ where $z_i > 0$ if $y_i = 1$ or $z_i < 0$ if $y_i = 0$.
2. Select model γ' with probability $q_{\zeta^{(t)}}(\gamma'|\gamma)$.
3. Jump to the model γ' with probability
$$\alpha(\gamma, \gamma') = \min\left\{1, \frac{\pi(\gamma')q(\gamma|\gamma')\pi(z|\gamma')}{\pi(\gamma)q(\gamma'|\gamma\pi(z|\gamma)}\right\}.$$
4. Compute
$$\zeta^{(t+1)} = \rho(\zeta^{(t)} + s^{(t)}(\alpha(\gamma, \gamma') - \bar{\tau})).$$
5. If γ' is accepted, draw a sample $\theta_{\gamma'} \sim \mathrm{N}((X_{\gamma'}^{\mathrm{T}} X_{\gamma'} + V_{\gamma'}^{-1})^{-1} X_{\gamma'}^{\mathrm{T}} z, (X_{\gamma'}^{\mathrm{T}} X_{\gamma'} + V_{\gamma'}^{-1})^{-1})$.

We think it is a safe conjecture that the results regarding ergodicity with the correct stationary distribution of the ARWM algorithm carry over to our model selection setting. In Section 5 we present empirical evidence which suggests that these results hold.

5 SIMULATION RESULTS

The performance of the adaptive MCMC algorithms is evaluated using two microarray data sets. These are the Arthritis data (Sha *et al*, 2003) and the Colon Tumour data (Alon *et al*, 1999). Adaptive versions of all algorithms in Section 2 were tested and are denoted as follows:

1. ADH-H : Adaptive Holmes and Held algorithm
2. ADAG-LA : Adaptive automatic generic sampler with Laplace approximation
3. ADAG-IWLS : Adaptive automatic generic sampler with Iterated Weighted Least Squares approximation
4. ADC-M : Adaptive Conditional Maximization.

Non-adaptive versions of the algorithms are indicated by dropping the first two letters "AD". All the adaptive MCMC samplers start with initial parameter value $\zeta_0 = 0.5$ and use the positive sequence of real numbers $\left\{s^{(t)} = \zeta_0/t : \ t \in \mathbb{N}\right\}$. The number of iterations was 2,000,000, the burn-in period 100,000 and the thinning 10 resulting in an MCMC sample size T of 190,000. Finally, we specify the value 0.3 as an optimal acceptance rate $\bar{\tau}$ because the optimum effective sample size of the MCMC algorithms that explore the model and parameter space of our problem is obtained when acceptance rates are between 0.25 and 0.4. Adopting $\bar{\tau} = 0.234$ instead makes very little difference to our results.

Table 1: The effective sample size ESS, the CPU time in seconds of the adaptive and non-adaptive algorithms with relative efficiencies of the non-adaptive algorithm over the adaptive algorithm for the Arthritis and Colon Tumour datasets

Arthritis

Method	ESS	CPU	R.E
H-H ($\zeta = 0$)	35144	9795	0.99
H-H ($\zeta = 0.25$)	35489	9880	1.00
ADH-H	34902	9668	
AG-LA ($\zeta = 0$)	58001	40135	0.75
AG-LA ($\zeta = 0.5$)	80701	40346	1.04
ADAG-LA	80853	41966	
AG-IWLS ($\zeta = 0$)	50942	9781	0.74
AG-IWLS ($\zeta = 0.5$)	68233	9870	0.98
ADAG-IWLS	69126	9822	
C-M ($\zeta = 0$)	55641	49173	0.95
C-M ($\zeta = 0.5$)	66801	49993	1.12
ADC-M	64243	53944	

Colon Tumour

Method	ESS	CPU	R.E
H-H ($\zeta = 0$)	32752	13968	0.99
ADH-H	32228	13587	
AG-LA ($\zeta = 0$)	47354	41355	0.97
AG-LA ($\zeta = 0.25$)	53653	42345	1.08
ADAG-LA	54777	45716	
AG-IWLS ($\zeta = 0$)	45494	14165	0.88
AG-IWLS ($\zeta = 0.25$)	51971	14246	1.00
ADAG-IWLS	51231	14066	
C-M ($\zeta = 0$)	39596	48159	0.94
C-M ($\zeta = 0.25$)	42370	48313	1.00
ADC-M	41861	47789	

The efficiency of an MCMC sampler can be measured using the Effective Sample Size (ESS) which is $T/(1+2\sum_{j=1}^{\infty}\rho_j)$ for an MCMC run of length T with lag j autocorrelation ρ_j (see *e.g.*, Liu, 2001). The interpretation is that the MCMC sampler has the same accuracy of estimates as a Monte Carlo sampler (where all the draws are independent) run for ESS iterations. In this paper, the ESS is estimated using the Initial Positive Sequence Estimator (Geyer, 1992). The algorithms have different running times and so we define the efficiency ratio for a sampler to be

$$\text{ER(Sampler)} = \frac{\text{ESS(Sampler)}}{\text{CPU(Sampler)}},$$

which standardizes the effective sample size by CPU run time and so penalizes computationally inefficient algorithms. We are interested in the performance of each adaptive algorithms to the non-adaptive algorithm with $\zeta = 0$ (which is the standard MCMC proposal for these types of models and represents a baseline) and with the optimal value of ζ among five candidates ($\zeta = 0$, 0.25, 0.5, 0.75, 0.95) which results in the highest ER. The relative efficiency of the non-adaptive over the adaptive algorithm is defined by

$$\text{R.E} = \frac{\text{ER(Non-Adaptive)}}{\text{ER(Adaptive)}}.$$

Table 1 presents result of the adaptive algorithms and various non-adaptive algoritms with fixed values of ζ for the Arthritis and Colon Tumour datasets. The relative efficiency for all sampling methods against the standard proposal ($\zeta = 0$) is always less than 1 indicating that the adaptive method is superior. The increase in performance depends on the particular method and the form of the posterior. However, the effect can be large in some cases. For example, the standard method only obtains 75% of the efficiency of the adaptive method with the Arthritis data and AG-LA and AG-IWLS algorithms. It is also clear that the Automatic Generic methods gain the most benefit. In fact, the effective sample size of the adaptive algorithms are very similar to the *optimal* non-adaptive algorithms in terms of mixing. Furthermore, the increase in CPU time of the adaptive algorithms is small. This leads to relative efficiencies quite close to 1 and therefore the adaptive algorithms achieve essentially the same efficiency as the optimal non-adaptive algorithms. Crucially, however, the adaptive algorithms avoid the pilot runs needed to tune the model proposal parameter ζ.

Figure 1 and Figure 2 show the trace plots of both the model proposal parameter ζ (left panels) and the empirical acceptance rate (right panels) of the adaptive algorithms for the Arthritis and Colon Tumour datasets, respectively. The parameter ζ of each adaptive algorithm converges to a value close to the optimal one obtained by manual tuning. Furthermore, the empirical acceptance rates converge to values quite close to the target acceptance rate 0.3. These results illustrate that the adaptive MCMC algorithms automatically find model proposal parameters ζ that give asymptotically the optimal acceptance rate $\bar{\tau} = 0.3$.

Figure 3 displays the scatter-plots of the log estimated posterior gene inclusion probabilities of the adaptive and optimal non-adaptive algorithms for the Arthritis and Colon Tumour datasets. The log posterior gene inclusion probabilities are very similar indicating empirically that the stationary distribution of the stochastic process generated by the adaptive MCMC algorithms is the target joint posterior distribution $\pi(\theta_\gamma, \gamma|y)$.

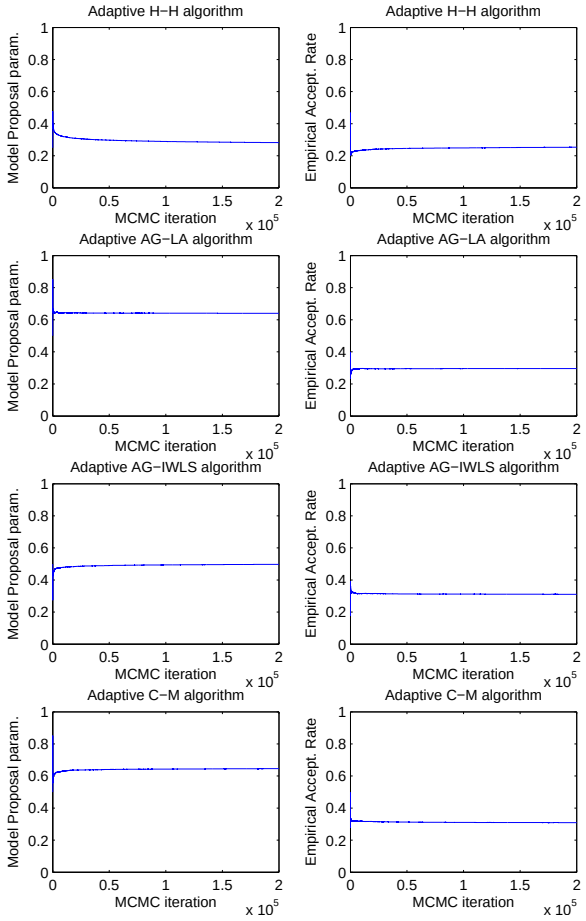

Figure 1: Trace plots of the model proposal parameter ζ and the empirical acceptance rate of the adaptive algorithms for the Arthritis dataset

6 DISCUSSION

This paper describes an adaptive Monte Carlo algorithm for posterior simulation for variable selection in probit regression models with many regressors. The algorithm leads to Markov chains with good mixing properties and without the need for pilot runs. In fact, the effective sample sizes for the adaptive algorithms are almost identical to those for the algorithms run at an optimized value of the proposal parameter ζ (found using trial-and-error). The methods are useful when there are a large number of variables that could potentially be included in the model, which leads to high average acceptance rates for standard algorithms. If the number of regressors is not large, then acceptance rates will not be high and an average acceptance rate of 0.3 may not be achievable. In this case, ζ should be close to zero and the value of $\zeta^{(t)}$ will converge to a value close to zero showing the robustness of the algorithm. Therefore, we suggest the use of adaptive MCMC algorithms to explore efficiently the model space of Bayesian variable selection in probit regression with many covariates. High acceptance rates for standard MCMC algorithms

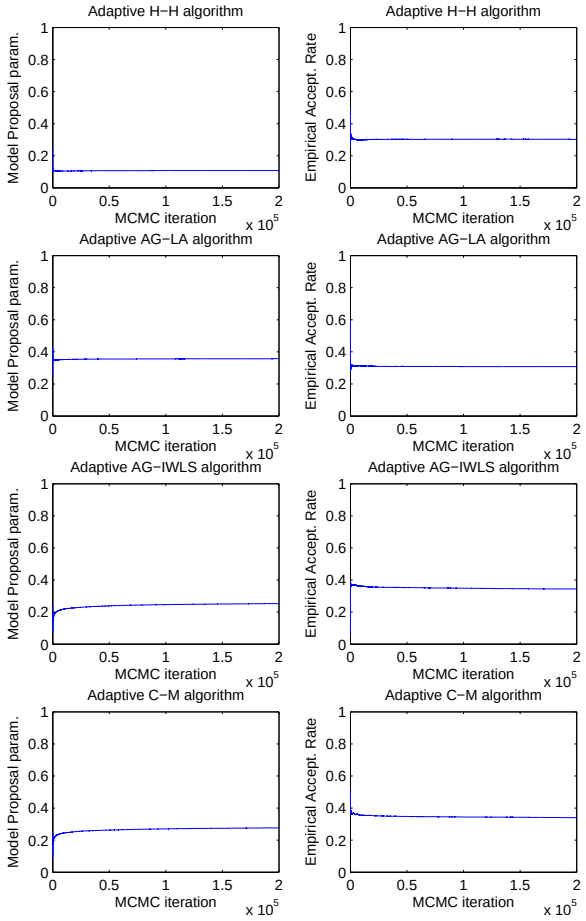

Figure 2: Trace plots of the model proposal parameter ζ and the empirical acceptance rate of the adaptive algorithms for the Colon Tumour dataset

in variable selection are also observed with other models such as linear regression when there are many regressors. The application of these adaptive methods to other models is an area of future research.

Acknowledgements

Demetris Lamnisos would like to acknowledge support from the Centre for Research in Statistical Methodology (CRiSM) at the University of Warwick.

References

J. H. Albert and S. Chib (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88**: 669-679.

U. Alon, N. Barkai, D. A. Notterman, K. Gish, D. Ybarra, D. Mack and A. Levine (1999). Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor

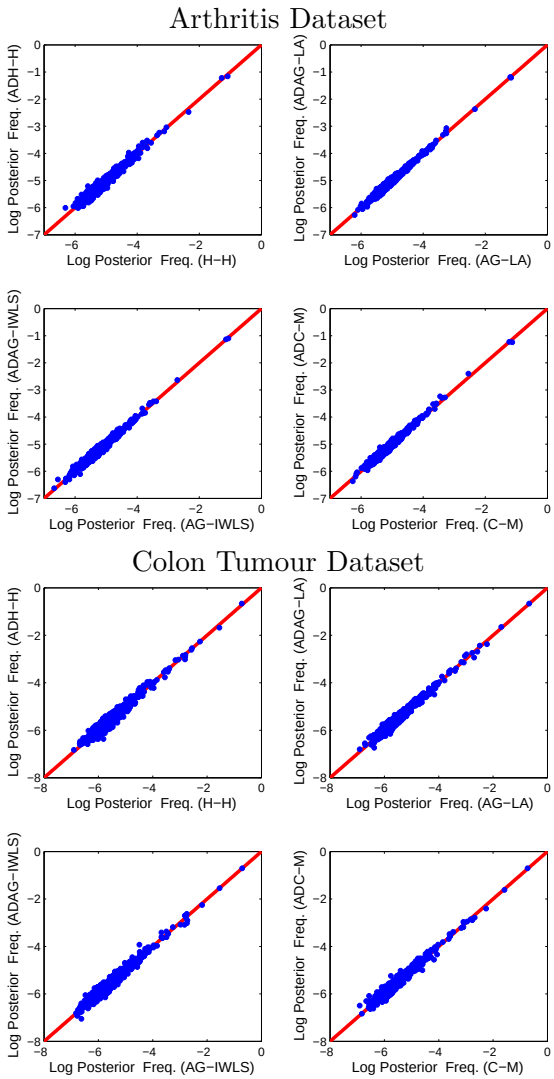

Figure 3: Scatter-plots of the log estimated posterior gene inclusion probabilities of the adaptive and optimal non-adaptive algorithms for the Arthritis and Colon Tumour datasets

and Normal Colon Tissues Probe by Oligonucleotide Arrays. *Proceedings of the National Academy of Sciences of the USA* **96**: 6745-6750.

C. Andrieu and J. Thoms (2008). A tutorial on adaptive MCMC. *Statistics and Computing* **18**: 343-373.

Y. F. Atchadé and J. S. Rosenthal (2005). On Adaptive Markov Chain Monte Carlo Algorithms. *Bernoulli* **11**: 815-828.

S. P. Brooks, P. Giudici and G. O. Roberts (2003). Efficient Construction of Reversible Jump Markov Chain Monte Carlo Proposal Distributions. *Journal of the Royal Statistical Society, Series B*, **65**, 3-55.

S. Dudoit, J. Fridlyand and T. P. Speed (2002). Comparison of Discrimination Methods for the Classification of Tumours Using Gene Expression Data. *Journal of the American Statistical Association*, **97**, 77-87.

C. J. Geyer (1992). Practical Markov Chain Monte Carlo. *Statistical Science*, **7**, 473-511.

P. J. Green (2003). Trans-Dimensional Markov chain Monte Carlo. In P. J. Green, N. L. Hjort and S. Richardson (eds), *Highly Structured Stochastic Systems*, 179-198. Oxford, U. K.: Oxford University Press.

H. Haario, E. Saksman and J. Tamminen (2001). An adaptive Metropolis algorithm. *Bernoulli* **7**: 223-242.

C. C. Holmes and L. Held (2006). Bayesian Auxiliary Variable Models for Binary and Multinomial Regression. *Bayesian Analysis* **1**: 145-168.

T. S. Jaakkola and M. I. Jordan (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing* **10**: 25-37.

D. S. Lamnisos, J. E. Griffin and M. F. J. Steel (2009). Transdimensional sampling algorithms for Bayesian variable selection in classification problems with many more variable than observations. *Journal of Computational and Graphical Statistics* **18**: 592-612.

K. E. Lee, N. Sha, R. Dougherty, M. Vannucci and B. K. Mallick (2003). Gene Selection: A Bayesian Variable Selection Approach. *Bioinformatics*, **19**, 90-97.

J. S. Liu (2001). *Monte Carlo Strategies in Scientific Computing*. New York: Springer-Verlag.

D. J. Nott and R. Kohn (2005). Adaptive sampling for Bayesian variable selection. *Biometrika* **92** 747-763.

Y. Qi and T. S. Jaakkola (2007). Parameter Expanded Variational Bayesian Methods. In B. Schölkopf, J. Platt, T. Hofmann (eds.), *Advances in Neural Information Processing Systems 19*, 1097-1104 . Cambridge: MIT Press

G. O. Roberts and J. S. Rosenthal (2001). Optimal Scaling of Various Metropolis-Hastings Algorithms. *Statistical Science* **16**, 351-367.

J. S. Rosenthal (2008). Markov chain Monte Carlo Algorithms: Theory and Practice. MCQMC'08 Conference Proceedings.

N. Sha, M. Vannucci, P. J. Brown, M. Trower and G. Amphlett (2003). Gene Selection in Arthritis Classification with Large-Scale Microarray Expression Profiles. *Comparative and Functional Genomics*, **4**, 171-181.

N. Sha, M. Vannucci, M. G. Tadesse, P. J. Brown, I. Dragoni, N. Davies, T. C. Roberts, A. Contestabile, M. Salmon, C. Buckley and F. Falciani (2004). Bayesian Variable Selection in Multinomial Probit Models to Identify Molecular Signatures of Disease Stage. *Biometrics*, **60**, 812-819.

C. Sherlock and G. O. Roberts (2009). Optimal scaling of the random walk Metropolis on elliptically symmetric unimodal targets. *Bernoulli*, **15**, 774-798.

K. Y. Yeung, R. E. Bumgarner and A. E. Raftery (2005). Bayesian Model Averaging: Development of an Improved Multi-Class Gene Selection and Classification Tool for Microarray Data. *Bioinformatics*, **21**, 2394-2402.